SEARCHING FOR SYMBOL STRING

FIELD OF THE INVENTION

[0001] The present invention relates to the search of an input symbol string among symbol strings. The invention is especially well suited for use in error-correcting database searches, in which the correct symbol string or the symbol strings closest to it are found in spite of an error in the input.

BACKGROUND OF THE INVENTION

[0002] Speech recognition, optical reading, correspondence searches of gene and protein sequences in bioinformatics, and database searches in general are examples of situations, in which there is a need to find a specific input symbol string among symbol strings. The symbol string can then be made up for example of consecutive characters or consecutive symbols representing phonemes. Often there is a danger that the input symbol string is not completely correct. The aim is, however, to find among the symbol strings of a database, for instance, the symbol string that completely corresponds to the input symbol string, or the symbol string that resembles it the most, if a fully corresponding input symbol string cannot be found.

[0003] A solution for searching for a symbol string is previously known, in which the symbol string is searched among symbol strings made into a trie data structure. The symbol strings are then grouped into branches in such a manner that all symbol strings starting with the same symbols belong to the same branch. The symbol strings in one branch divide into new branches at the symbols, from which onwards the symbol strings differ from each other.

[0004] The "tree-like" trie data structure has been employed in the search for symbol strings in such a manner that the branches of a data structure are searched until the leaves. Each new symbol encountered on the branch indicates a calculation point, at which a distance is calculated between a sample symbol string formed by the symbols of the calculation point and the calculation points preceding it and the searched input symbol string by comparing them in alternative ways. The distance refers to any reference value that describes how many changes are required to make the compared symbol strings correspond to each other. One known way of calculating the distance is the Levenshtein algorithm.

[0005] The calculation ends when the distances for all calculation points of all branches of the trie data structure are calculated. After this, a

comparison is made to find the shortest distance. To produce a response, the symbol string of the branch or the symbol strings of the branches with the shortest distances in the last calculation points are selected.

[0006] The most significant weakness of the above-mentioned prior-art solution is that it requires a relatively large amount of calculation. The best possible symbol string, i.e. the one closest to the input symbol string, can only be found after all calculation points in the trie data structure are calculated. Because in database searches, for instance, the number of symbol strings in the database is extremely large, this means that the number of required calculations becomes very large and, therefore, the time required for the calculations is long. Obtaining a response to the input, therefore, requires a lot of time.

BRIEF DESCRIPTION OF THE INVENTION

[0007] It is an object of the present invention to solve the above-mentioned problem and to provide a solution that makes it possible to reduce the number of calculations required to produce a response, thus making the production of the response faster than before. This object is achieved by the method of independent claim 1, the computer program of independent claim 5, the data medium of independent claim 6, and the apparatus of independent claim 7.

[0008] The solution of the invention is based on the idea that the number of calculations required to search for the symbol string and, thus, to produce a response can be significantly reduced, when for each distance, the shortest possible length difference corresponding to it is also calculated, as well as a reference value on the basis of the distance and the length difference. Said reference value then indicates the best possible distance that can theoretically be achieved when proceeding to the end of the branch in question, upon the condition that all the symbols remaining on the branch correspond to the unexamined symbols of the input. In such a situation, the deciding factor is the length differences between the input and the symbol strings. When the input and symbol string are of different length, each "extra" symbol increases the distance between them. Because it is possible to determine the best possible reference value at each calculation point, it is also possible to determine by comparing the reference values, which of the branches may provide the shortest possible distance. In such a case, only the branches in ques-

tion are examined and calculation is skipped on the branches whose reference value indicates that a better distance than in the other branches cannot be achieved in them.

[0009] Due to the solution of the invention, calculation can be skipped for a large part of the calculation points on the branches of the trie data structure without endangering the finding of the best symbol string. This, in turn, reduces significantly the time required for calculation, and the search for the best symbol string or symbol strings is faster than before.

[0010] In one preferred embodiment of the invention, the distance of the symbol string (or symbol strings) used in producing the response and the input symbol string are compared with a predefined maximum distance, i.e. limit value. If the distance exceeds the maximum distance, this means that the found symbol string differs so much from the input symbol string that forwarding it in a response is not expedient (a symbol string that sufficiently resembles the input symbol string has not been found). The produced response is then altered before it is transmitted to indicate that the input symbol string was not found.

[0011] In a second preferred embodiment of the invention, said lowest reference value is compared during branch selection with a predefined maximum distance, and the calculation is ended if the lowest reference value exceeds the maximum distance. The reference value represents the best possible obtainable distance, if the rest of the symbols on the branch correspond to the symbols left in the input and the numbers of symbols match. If, under the circumstances, the lowest reference value exceeds the maximum distance, it means that the input symbol string or one resembling it will not be found among the symbol strings, and the calculation and, at the same time, the search for the input symbol string can be ended as unnecessary.

[0012] A third preferred embodiment of the invention checks during branch selection, whether calculation has already been done for the last calculation point on one of the branches, and ends the calculation, if it turns out that for the last calculation point on a branch, a reference value has been obtained that is lower than the reference values obtained from all the other calculation points. This way, the calculation and symbol string search can be ended already before calculating all calculation points on all branches, because due to the use of the reference values, it has been established that the symbol string of a branch calculated to the end corresponds best to the input symbol string.

[0013] Preferred embodiments of the method of the invention are set forth in the attached dependent claims 2 to 4.

BRIEF DESCRIPTION OF THE FIGURES

[0014] In the following, the invention will be described by way of example in greater detail with reference to the attached figures, in which

Figure 1 is a flow chart of the first preferred embodiment of the method of the invention,

Figures 2a to 2f illustrate the progress of the search for the input symbol string when following the flow chart of Figure 1, and

Figure 3 is a block diagram of the first preferred embodiment of the apparatus of the invention.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0015] In the following, the invention will be described by way of example in greater detail with reference to the flow chart of Figure 1 and the calculation example of Figures 2a to 2f.

[0016] In block A of Figure 1, a trie data structure is created in a manner known per se and the symbol strings are grouped into branches of the trie data structure. Figure 2a shows this type of data structure with grouped symbol strings: ABACUS, ABOARD, BOARD, and BORDER. The creation of a trie data structure is a step that does not necessarily need to be repeated each time a new symbol string is searched from the database. The earlier created data structure can be utilized for instance as long as no new symbol strings are added to the database. That is, only when a new symbol string is added to the database, will the creation of a new trie data structure be necessary.

[0017] In block B of Figure 1, an input symbol string, i.e. the symbol string that will be searched for among the set of symbol strings, is received. The next description assumes by way of example that the searched input symbol string is ABORD.

[0018] In block C, the routine starts by proceeding to the first calculation point P, at which according to block D, the distance between the input and a sample symbol string formed by the symbols of the calculation point of the branch in question and the calculation points preceding it is calculated in a manner known per se by using the Levenshtein algorithm. In addition, according to the invention, length differences D and reference values R are calculated as follows. For calculation point P, the sample symbol string is " " (empty char-

acter in the beginning) and the input symbol string is (with the empty character added to the beginning) " ABORD". The calculation proceeds as follows.

[0019] - The distance between empty characters " " is 0, because the characters match each other. There are 5 characters left in the input symbol string, and 5 to 6 characters left in the length of the symbol strings passing through point P, which is marked at point P as max 6, min 5. The shortest possible length difference is thus 5-5=0. The reference value is distance + length difference, i.e. 0+0=0.

[0020] - The distance between the symbol strings " " and " A" is 1, because the symbol strings being compared match each other by adding/changing one character. There are 4 more characters left in the input symbol string, and 5 to 6 characters left in the length of the symbol strings passing through point P. The shortest possible length difference is thus 5-4=1. The reference value is 1+1=2.

[0021] - The distance between the symbol strings " " and " AB" is 2, because the symbol strings being compared match each other by adding/changing two characters. There are 3 more characters left in the input symbol string, and 5 to 6 characters left in the length of the symbol strings passing through point P. The shortest possible length difference is thus 5-3=2. The reference value is 2+2=4.

[0022] - The distance between the symbol strings " " and " ABO" is 3, because the symbol strings being compared match each other by adding/changing three characters. There are 2 more characters left in the input symbol string, and 5 to 6 characters left in the length of the symbol strings passing through point P. The shortest possible length difference is thus 5-2=3. The reference value is 3+3=6.

[0023] When the compared symbol strings have been compared in all alternative ways, the results shown in the table at the bottom of Figure 2a are obtained.

[0024] In block E, the calculation point is searched that has provided the lowest reference value. The lowest reference value has been obtained for point P, which is the only calculation point calculated so far. Therefore, the routine proceeds from this calculation point along branch 1.

[0025] In block F, it is checked whether the condition to terminate the calculation is fulfilled. There may be several termination conditions. The following mentions by way of example two termination conditions.

[0026] Termination condition 1: The calculation is terminated, if the lowest reference value exceeds a predefined maximum distance. In such a case, the conclusion is that the searched input symbol string differs so much from the set of symbol strings that the search can be interrupted, because a symbol string resembling the input symbol string will not be found. The definition of a suitable maximum distance depends on the application. This example assumes that the maximum distance is 5.

[0027] Termination condition 2: The calculation is terminated, if, on a branch, the calculation has already been done for the last calculation point, and for the last calculation point, a reference value has been obtained that is lower than the reference values obtained for all the other calculation points. Therefore, the calculation and search for symbol string can be terminated already before the calculation is finished at the calculation points of all branches, because due to the use of the reference values, it has been established that the symbol string of a branch calculated to the end corresponds best to the input symbol string.

[0028] The table in Figure 2a shows that the lowest reference value R is 0 that is smaller than the maximum distance 5. Therefore, the first termination condition is not fulfilled. The second termination condition is also not fulfilled, because the last calculation point has not yet been reached in any of the branches. Therefore, in the block diagram of Figure 1, the routine enters block C, and in Figure 2a, the routine proceeds along branch 1 to point P1.

[0029] The calculation according to Block D is repeated in point P1. The sample symbol string is " A" and the input symbol string is still " ABORD". To facilitate the calculation of the distances of point P1, the calculation is started using the distance calculations done in point P that are transferred to the table at the bottom of Figure 2b. The calculation proceeds as follows.

[0030] - The distance between the symbol strings " A" and " " is 1, because the symbol strings being compared match each other by adding/changing one character. There are 5 more characters left in the input symbol string, and 5 characters left in the length of the symbol strings passing through point P1, which is marked at point P1 as max 5 min 5. The shortest possible length difference D is thus 5-5=0. The reference value R is distance + length difference, i.e. 1+0=1.

[0031] - The distance between the symbol strings " A" and " A" is 0, the symbol strings match each other. There are 4 more characters left in the

input symbol string, and 5 characters left in the length of the symbol strings passing through point P1. The shortest possible length difference is thus 5-4=1. The reference value is 0+1=1.

[0032] - The distance between the symbol strings " A" and " AB" is 1, because the symbol strings being compared match each other by adding/changing one character. There are 2 more characters left in the input symbol string, and 3 characters left in the length of the symbol strings passing through point P1. The shortest possible length difference is thus 5-3=2. The reference value is 1+2=3.

[0033] The distance between the symbol strings " A" and " ABO" is 2, because the symbol strings being compared match each other by adding/changing two characters. There are 2 more characters left in the input symbol string, and 5 characters left in the length of the symbol strings passing through point P. The shortest possible length difference is thus 5-2=3. The reference value is 2+3=5.

[0034] When the compared symbol strings have been compared in all alternative ways, the results shown in the table at the bottom of Figure 2b are obtained. It should be noted, however, that the above method for calculating distances between compared symbol strings is only one example, and in addition to it, there are other known and possibly even simpler methods. It is not essential for the invention how the distances are calculated. One alternative for calculating distances is to utilize a table of the type shown at the bottom of Figure 2b, and especially the preceding calculated distance column.

[0035] When the calculation of point P1 is done, the calculation point with the lowest reference value R is again searched for in block E. The result is point P whose reference value is 0 that is lower than the lowest reference value 1 of point P1. Therefore, the routine proceeds next along branch 2 to point P2. In block F, it is detected that the termination conditions are not fulfilled, after which the routine enters block C to repeat the calculations for point P2.

[0036] In the following, the calculations of all calculation points are not examined, but the routine moves directly to the situation shown in Figure 2c, in which the calculations are done for points P3 and P4. In block E of the block diagram of Figure 1, it is then found that the lowest reference value R is obtained at calculation point P3, the lowest reference value R of which is 1 in

Figure 2c, whereas the lowest reference value R at point P4 is 2. Therefore, the routine proceeds next along the branch of calculation point P3.

[0037] Figure 2d shows a situation, in which the calculations of calculation points P5 and P6 are done. In block E of the block diagram of Figure 1, it is then found that the lowest reference value R is obtained at two calculation points, i.e. the lowest reference value of both calculation point P5 and P6 is 1. Next, the routine follows the branch of calculation point P5.

[0038] Figure 2e shows a situation, in which the calculations of calculation point P7 are done. In block E of the block diagram of Figure 1, it is then found that the lowest reference value R is obtained at calculation point P6, at which the lowest reference value R is 1 (the lowest reference value R of calculation point P7 is 2). Next, the routine follows the branch of calculation point P6.

[0039] The figures do not show all intermediate steps, but when the calculations are repeated at calculation points P9 and P10, the reference value R is 1 for these points. When the calculations are again repeated at calculation point P10, the situation is as shown in Figure 2f, when block E of the block diagram of Figure 1 is again reached. The lowest reference value R of calculation point P10 is 1. Because the reference values of the calculation point calculated last in all the other branches (on the branch, to which point P7 belongs, R=2, and on the branch, to which point P4 belongs, R=2) are higher than the reference value R=1 of calculation point P10, it is calculation point P10, from which the routine should proceed next. However, calculation point P10 is the last calculation point on the branch. Therefore, in block F, it is found that the termination condition 2 described above is fulfilled, and calculation can be terminated.

[0040] In block G, the symbol string of the branch that led to calculation point P10 is selected for producing the response. The symbol string in question is ABOARD. This symbol string is provided as response to the input.

[0041] Differing from the block diagram of Figure 1, it is possible to make an extra check after block G. Then, the distance of the symbol string (or symbol strings) used to produce the response and that of the input symbol string are compared with a predefined maximum distance, i.e. limit value. In the situation of Figure 2f, the distance between the symbol string ABOARD used to produce the response and the input symbol string ABORD is 1 (circled in Figure 2f). If the distance exceeds the maximum distance, it means that the

found symbol string differs so much from the input symbol string that transmitting it on in the response is not expedient (a close enough symbol string has not been found). The produced response is then changed before it is transmitted on to indicate that the input symbol string was not found. This way, it is possible to avoid a situation, in which the response becomes a symbol string that is very much different from the input symbol string.

[0042] Figure 3 is a block diagram of the first preferred embodiment of the apparatus of the invention. In Figure 3, the apparatus 10 is illustrated using functional blocks 11 to 18. However, it is important to note that the actual structure of the apparatus may differ from what is shown in Figure 3. The functions of the blocks in Figure 3 can in practice be implemented by one or more circuits or computer programs, or alternatively by a combination of circuits and programs. It is then also possible that the functions of the apparatus are not implemented exactly as illustrated, but the functions of one or more blocks can be combined in one circuit or program.

[0043] The apparatus 10, by means of which the method described in Figures 1 and 2a to 2f can be used, can be a computer connected to a telecommunications network and containing a memory 13 with a database of symbol strings in it. The apparatus comprises means 12 for creating a trie data structure by grouping the symbol strings stored in the memory 13 into branches of the trie data structure.

[0044] When the trie data structure is created and an input symbol string received through an input 11 of the apparatus, the apparatus 10 begins to search for the symbol string that best corresponds to the input symbol string in the memory 13. To do this, the apparatus comprises calculation means 14 for calculating distances, length differences and reference values between a sample symbol string formed by the calculation point and the calculation points preceding it in the examined branch and the input symbol string by comparing these in alternative ways.

[0045] The apparatus also has selection means 15 that repeatedly select the next branch, along which to proceed, and indicate to the calculation means 14 the next calculation point for calculation, as earlier described in connection with the flow chart of Figure 1. When the selection means 15 detect that a termination condition is fulfilled, i.e. that the calculation should be terminated, they inform means 16 of this.

[0046] The means 16 select on the basis of the information in the memory 13 one or more symbol strings, the distance of which to the input symbol string is the shortest on the basis of the performed calculations. After this, production means 17 produce and transmit through an output 18 of the apparatus 10 a response, which is thus made up of the symbol string or symbol strings that most resemble the input symbol string.

[0047] It is to be understood that the above description and the related figures are only intended to illustrate the present invention. It will be apparent to a person skilled in the art that various modifications can be made to the invention without departing from the scope of the invention disclosed in the attached claims.